DC05\_transcript\_deidentified

**SUMMARY KEYWORDS**

data, qualitative data, worked, qualitative research, curating, metadata, release, study, identify, curation, question, restricted access, qualitative study, documentation, context, people, mask, transcripts

**SPEAKERS**

Sara Mannheimer, DC05

**Sara Mannheimer** 00:01

Great. So the idea behind my research is that basically big social research, big social data, like social media data, and qualitative data that's being reused have like similar ideas behind them. They're both sort of like narrative data that are being used for a different purpose than they were originally created for. But those similarities haven't really been investigated in research yet. And so I'm wondering about curation practices that have been developed for qualitative research, which qualitative data reuse has been going on for several decades, you know, whereas big social data reuse is still a little bit more nebulous. So I'm hoping that by connecting these two communities and talking to data curators who have worked with one or both of the communities will be able to learn more about how data curation can support responsible research for big social and qualitative data reuse. And so through a lit review identified six issues that are challenges for both groups. And so the interview is structured around the six issues, which are context data quality data comparability informed consent, privacy and confidentiality and intellectual property. So I have basically six questions. There's like some intro-, introductory questions and a wrap up question, but one question for each of those issues. All right, so tell me about you. Your work is what types of data you usually curate, and then what your interests are around data curation.

**DC05** 01:52

Okay, well, I'm a data curator at [organization]. I started there [date redacted—DC05 is in their early career]. Before that, I had been a graduate student at [university], in [a social science department], so I actually used [archived] data in my work. And so I thought it would be kind of cool to once I graduated to get a job [as a data curator], and actually get the behind the scenes, look at how these datasets that I've been using, are, you know, the behind the scenes stuff. So when I got [to this organization], I knew going in that mostly it was going to be quantitative data sets. But I actually have a qualitative research background. So when I had heard that, we did get some studies with qualitative, I was like, "Oh, I told my supervisor, you know, like, I have experience in this area. So if you ever get any qualitative studies, I would be happy to take them on." And she said, "That's great, because, you know, there aren't a lot of people in the [specific part of the organization] that have been that have experience with qualitative data. So that's great."

**Sara Mannheimer** 03:23

Awesome. Um, and so did you, have you already identified a recent time when you curated a qualitative data set? I'm trying to like come up with a specific time just to help us get more specifics as we talk to you to question in order to stick just to that that time, but I think it would be helpful to have something in mind.

**DC05** 03:46

Yeah, I'm actually working on one right now. Perfect, perfect timing.

**Sara Mannheimer** 03:52

Cool. That'd be great. Because you'll have it fresh in your mind. So what is your exam-, tell me about your example.

**DC05** 04:01

Okay, so the qualitative study that I'm working on curating right now was part of a larger quantitative study [about families and community programs]. So it's a very large study, it's [hundreds of] transcripts, so [thousands of] pages of text. So the kind of the largest qualitative study I've worked on yet. I just finished one that was significantly less than that it only had like [less than 10] transcripts in it.

**Sara Mannheimer** 05:21

Wow. So do you, are you reading through each individual page and curating it in that way? Or is it more of a scan?

**DC05** 05:29

Um, so about this study, the, the research team had gone through and de-identified, they de-identified like location, names, and any rare or publicized events that can be used to identify participants. So we got that information when we got the data, and I'm like, okay, um, but having gone through and reviewed other qualitative deposits through, that we've gotten from other PIs, sometimes they don't catch everything. So, we'd still do an in depth review of the data. Because of the volume of data, we usually [review some of the dataset] when we get it in it for just at least for [as an initial review]. So whenever we get any kind of data quantitative or qualitative part of our process is to look at the variables or look at the text and make sure that there's no direct identifiers or personally identifying information in it. And then we have strategies like masking names, masking locations, to try and reduce [the risk of the data being identifiable]. So I was looking through about [a fraction of] the entire dataset for that initial review, but I will eventually be reading the entire transcripts to catch anything I missed in that initial review.

**Sara Mannheimer** 07:10

Wow, an amazing service. Do you know if the if the data from this study was part of a grant funded, was grant funded, and like required specific sharing or treatment of the data for that reason?

**DC05** 07:27

Yes, it was, um, most of the projects that I work on are, they're grant funded, somehow, let me look up the actual PI information, [names agency].

**Sara Mannheimer** 07:52

Okay. So did you have like a data management plan in with the data? Or do you know anything about that?

**DC05** 08:02

Not for this particular one, the only information we got about how the data were managed was in the user manual and what they had masked... to kind of zip to a different example. I also work on [government agency] studies, and which comes with a lot of mixed methods data. And they are required to submit like a data management plan or a data, like, "What is your plan for how this data will be stored?" So I have experienced with those and for this one that I'm currently working on, I did not get one.

**Sara Mannheimer** 08:39

Okay, cool. So just as a side question, when you get those DMPs from [government agency] studies, does that inform like, how you curate the data? Do you go through and see what they said they would do and try and align with it? Or do you still sort of do what you think is best from [your organization's] point of view?

**DC05** 08:57

Um, it's a little bit of both. So when they give us a document that says, here's what we did to de-identify the data, that's very helpful, because then it gives us an idea of, you know, what they were looking for, and if they had missed anything, it helps us kind of align with like, say, like, even with my particular example, they had said that they've masked any geographic locations. And when I'm reading through the data, if there is like a name of an organization or a landmark that could tip off the location of the site, I can say okay, they probably missed this one and they would have marked it if they recognized it, so I would feel okay and like going in and masking it if they missed it.

**Sara Mannheimer** 09:52

And then what are your plans? What are [your organization's] plans for storing, retaining, deleting the data in the future is it kept in perpetuity?

**DC05** 10:02

I don't have a lot of information on that, either quantitative or qualitative. I mean, I know we've got studies going back to like, you know, data going back [decades]. But I think qualitative data is like a newer thing. I'm not exactly sure.

**Sara Mannheimer** 10:24

Okay. And is it an open access collection or restricted? And we'll talk more about that further down.

**DC05** 10:32

Most of the time, qualitative data is restricted access. So if somebody wants to get a hold of any restricted access data, depending on [where it is housed], they'll normally need to go through a process where they have to apply for restricted data use. And it usually requires sending in an application with an IRB protocol.

**Sara Mannheimer** 11:00

Okay. All right. So, now we'll get to like the meat of the interview with six issues. So the first issue is context. And since we're talking about qualitative data, I'll read you this first paragraph. "Qualitative research is a process that may include deep and prolonged contact and connection with research subjects, trying to understand subjects within their own context. And so qualitative data are highly context dependent, as Hines et al. wrote, context is a source of data meaning and understanding ignoring context, under using it or not recognizing one's own context driven perspective, will result in incomplete or missed meaning and understanding of human phenomena and misunderstanding." So that's kind of to get you situated in where I'm coming from in terms of context. And so in your example, what challenges did you encounter, if any, when you are trying to document or capture the context in which the data was collected?

**DC05** 12:18

Because these... because the sample is [from a broad geographical area], and because the locations are masked, you lose the context of the particular site itself. So if staff members [from the community programs studied] have like a promising program, or techniques, or if they have a particular community makeup that allows them to be successful, like, you don't really glean that from the interview. So I feel like that context is kind of that context is missing. So and we try to capture, like, as much as we can in by putting it in metadata. But sometimes we just a lot of the time, we don't get that like, extra information about like, you know, locale which is very important to qualitative.

**Sara Mannheimer** 13:19

Do you ever reach back out to interview or researchers, as part of your curation work, to like ask for more of that? Or is it generally you take what you've received, and then do what you can in terms of adding metadata?

**DC05** 13:35

It really depends on how responsive the PIs are. Sometimes, we will try to reach back out if we feel we need more information. A lot of the studies that I've worked on the PIs do a good job of including any relevant documentation upfront. And for [federally funded] projects, some of that documentation is required. So like getting the IRB protocol and getting the final report that they're required to submit, that is a wealth of information that we can use to input into our metadata and explain to end users what you know, the context is and how the study was designed and everything. So we can try to reach back out to the PI. But sometimes, especially if it's been an arduous process to get even basic information about it. It's it's much more of a struggle.

**Sara Mannheimer** 14:39

Yeah. Okay, so your main strategies for documenting context are trying to get as much like supplementary documentation from the researchers as possible? And then include those as metadata is that right? Anything else I'm missing?

**DC05** 15:02

Not that I can I think that's pretty much, you've got the core of it.

**Sara Mannheimer** 15:10

Okay, we're going to skip these sub-questions since, have you, like, helped researchers do big social research or worked on archiving social media or other big social data?

**DC05** 15:22

No, I don't have any experience with big data.

**Sara Mannheimer** 15:24

Okay, we'll skip questions 1a, 2a etc. Okay, so data quality. In your example, what challenges did you encounter, if any, when you were trying to document data quality? So things like missing data, or the quality of the method? Or sampling and sort of methodology stuff? And then what strategies... If you did encounter challenges, what strategies did you use to try and help those challenges?

**DC05** 16:00

Um, this study was actually pretty easy going in terms of data, like their data. There wasn't really an issue with data quality that I've seen yet. The user guide that they gave us had a ton of information on how to link the qualitative transcripts with previous data as part of the larger study, which that was like my main concern, especially from a [reidentification] perspective, that these data can be linked to other data that we've already [made available]. So that they had that information upfront, I was like, okay, great. Now I know exactly how to tell users how to get to this other data, if they want to link the two data sets. Um, I mean, it was very descript-, like the documentation they provided was very descriptive. They explained file naming structures for all of the data files, the file count was correct. You know, they didn't say in the report, like, "Oh, we have 100 datasets," and then instead, I got 300. No, this, the numbers were consistent. And they explained the sampling. They, like I said, they explained how they de-identified it and quality control procedures. So yeah, it's really nice to get all that.

**Sara Mannheimer** 17:27

Yeah. How is that communicated to the users who would find the data in [the repository]? Is all that included, just as other documents that are part of the data set?

**DC05** 17:38

Yes. So we will [provide]... We normally [provide] any documentation, like for public facing, like any user guides, we also release the questionnaires. Less applicable with qualitative, but if PIs give us code books, like their own version of a code book we'll [make those available] as well. It's especially important with restricted data, because that is the only thing that users will see besides the public metadata. So we have to try to describe the data as best as we can. For, if someone is trying to decide whether or not they want to go forth through the process of actually obtaining that data through restricted access.

**Sara Mannheimer** 18:31

Okay, so you make most of the documentation public, even if it's restricted?

**DC05** 18:40

Yes.

**Sara Mannheimer** 18:40

Okay.

**DC05** 18:40

And we will also, depending on what is in the documentation, we'll also [review that] as well. Because there have been times where there have been either researcher names or other identifying information left in the documentation, so we'll make you know, will [redact and mask] that as well.

**Sara Mannheimer** 19:00

Okay. Nice. Well, I have a question about privacy, but and maybe we'll talk about it later. We'll talk about it when we get down there. Okay, so data comparability, did you encounter any challenges relating to comparability or interoperability of your data set? You did talk about linking the dataset with related data. Did you encounter any other challenges like metadata interoperability, or? Yeah, people try, or in general, people trying to use like multiple, trying to combine multiple data sets and challenges there.

**DC05** 19:47

In terms of interoperability, I don't foresee an issue with this data set. The directions are very clear on what, what ID numbers to link and the ID numbers are provided in the transcripts to link to the quantitative data. In terms of actually opening the data, like the file formats they received it in, I didn't have an issue with that because it was in Microsoft Word format. And then when we release this data set for public for, to, to the archive, it'll be in three different file formats. And we've talked about like, should we revisit the type of file formats that we released this in. So we'll release it in a PDF form, a Rich Text Format form and a plain text form.

**Sara Mannheimer** 20:46

Nice. So each document has each of those three formats?

**DC05** 20:53

Yes.

**Sara Mannheimer** 20:54

Okay. Um, how about for the metadata that you use [at your organization]? Do you have a standard you have, do you use DDI?

**DC05** 21:04

Yes, yeah.

**Sara Mannheimer** 21:06

Any other standardized metadata schemas that you use? Or do you have thoughts on? how, you know, metadata interoperability?

**DC05** 21:17

Um, I have less experience with that. So I...

**Sara Mannheimer** 21:22

That's okay. No worries. Just wanted to drill in there if you had thoughts. Okay, let's talk about informed consent. Did you encounter challenges relating to informed consent of the participant, particularly consent for future use of the data? And then what strategies did you use to address any challenges, you encountered.

**DC05** 21:49

No challenges on this one, we got the informed consent sheet as part of like the data deposit package. And, like for [government agency data], that's also required, it's a required document to make sure that, you know, everything is on the up and up in terms of informed consent, and that there's a statement in there it says, like these data will be deposited at, you know, [repository]. Some of the things that I do wonder is if, like, for the studies that I've worked on, just because I've done qualitative research in the past, is, especially for like interviews, and if it's part of a mixed methods project, like the the quality, like it explained in the informed consent, that qualitative data will also be archived. Because I'm not sure if that is something that that's taken into consideration, or...

**Sara Mannheimer** 23:00

Yeah, this is a... I mean, all of these issues are like not not solvable. But part of you know, like, part of my question is, like, is it important that the participants, they can consent to this, but can it ever really be informed? Because you don't exactly know, like, what future use your data data will be people will use your data for?

**DC05** 23:29

Right.

**Sara Mannheimer** 23:30

So, but...

**DC05** 23:36

Yeah, I do think that it's, especially for qualitative data, it's important to say, you know, at least explain that, you know, these are going to, you know, I'm going to put these transcripts on this website, and there's going to be like a bit of a, you know, a process to go through for somebody to obtain it. And we're going to de-identify the data. But, you know, there's, we have, you know, we have a level of trust in the people that are trying to access this data, hence the process, but there's no telling what they're going to do with it when they get it.

**Sara Mannheimer** 24:07

Yeah. I mean, I guess, with requiring an IRB, like approval showing the IRB approval does sort of guarantee that the people who are using it have certain ethical structures that they're following.

**DC05** 24:26

Yeah, and it's actually says in a lot of our restrictions, use language that, you know, these data are not to be used to try and identify people in the data.

**Sara Mannheimer** 24:38

Okay, so you'll have like a Terms of Use, basically.

**DC05** 24:43

Yes, yes.

**Sara Mannheimer** 24:44

Okay. Is that present for all the datasets or just the restricted ones?

**DC05** 24:50

Yeah, so we have a terms of use for every single data set that [is available]. It's in all of our documentation. It's in when users will sign up for an account, which allows them to gain access to publicly available data sets, they have to agree to a Terms of Use every time they try and download either documentation or data that's publicly available.

**Sara Mannheimer** 25:23

Have you... This is just out of curiosity, do you know if [your organization] has like worked with users to see how well people understand the terms of use? Or what's like the sort of enforcement of the terms? Do you know?

**DC05** 25:40

I'm not sure about enforcement, I do know that we have like a website team. And with a couple of user experience user designers that will look at like, look at the website and see if they can make it more user friendly. But as far as like understanding the terms of use, I don't know.

**Sara Mannheimer** 26:04

Okay, cool. All right, let's move to privacy and confidentiality. During your example, what challenges did you encounter relating to privacy for the people represented in the data? And then what strategies do you use to address the challenge? So you've talked about [redacting and masking identifying information]. And the restricted access. Tell me more about your procedures around privacy.

**DC05** 26:43

Sure, absolutely. So with every data set that we get quantitative, qualitative, whichever, we do [look for potential risk of reidentification]. So we will look at whether it's variables or text, and we have a guide like an internal guide as to what to look for, like what is the direct identifier versus an indirect identifier. At different [levels of access] we have, what can be left in the data versus what we need to mask or take out of the data or redact from the data? Obviously, not obviously, but direct identifiers are almost always taken out with the exception of the data that goes into our physical enclave, which means that a person has to physically come to [the location of the organization] to access that data. So that would be the only case where direct identifiers will be left in the data. Most, for anything else, like public use, or restricted use direct identifiers are taken out. Indirect identifiers are going to depend on the study itself, how it was conducted, like we have, like a [list] of questions that we ask like, "What is the sample size? Is the sample representative of few people in the study universe? What types of sensitive information is there? Does the study involve minors? Does the study involve other vulnerable populations? Can this data be linked to other people? Is there information on other people that weren't the respondent?" So we have like this little, so we put in that information and then we also judge, you know, how harmful would it be to the participants if this data were to be, you know, breached? And then how difficult or easy is it to identify somebody, so we kind of can take those considerations. And then we have like a matrix based on harm versus... harm and de-identification as to the recommendations we would make to de-identify the data further if it needs it. If we need to bump it to, let's say [a higher level manager], whoever was in charge of managing that data or obtaining that data, who suggested that it's okay for anybody, [any registered user] to obtain it. But we see something that's like, "Oh, this is you know, we need to bump this up to a restricted access." So we can make those recommendations to the [higher level manager] and normally, there might be some conversation, but normally, they'll tend to go with what [the curator] is suggesting, like, "Yeah, okay, let's bump this up to a restricted access."

**Sara Mannheimer** 29:51

And the [higher level manager is employed by your organization] too?

**DC05** 29:54

Yes, yes. And then just from my experience curating qualitative data, finding information that that could be identifiable or could identify people. There's a lot of internet searching of unfamiliar terms and like organization names and other things that pop up in a transcript and seeing if it links to a person or a location. So I'll give you an example for the current data set I'm working on. The transcripts don't have any geographic information in them, because based on what the PI said, they masked everything. You know, we just know that the site is located in [a specific country]. But the respondent in that trans-, in one of the transcripts talked about... they gave an organization name. And so I decided to just type it in on Google and using a, you know, you know, an incognito browser. And so, the organization name came up and then it tied, it was tied to a very specific city, and I'm like, oh, might need to mask that information, like the name of the organization, if, especially if they're wanting to mask all of their geographic information. So that's, you know, that's a big component of [curating] qualitative is just typing stuff up in Google and seeing if it can identify people.

**Sara Mannheimer** 31:43

Wow. That is a lot. I guess I'm curious about like, the time it takes to do that. And like, I don't know, I'm sort of coming up with this question on the fly. But like, what, what are your, like, your coworkers, thoughts on like, the time that you take to do all of this and the benefit, you know, like, do you would you take more time? If the if you know, when you went through this worksheet you saw there were more sensitive factors more harm, more potential harm? like would you then take more time to do more Googling or to be more detailed as you go through? Or is it like each time the process is super detailed, and you're spending the same amount of time on each dataset?

**DC05** 32:38

Well, we have different curation levels, depending on how intensive the curation process is going to be. Qualitative Data always is a [highest level] curation for us. So we know that we're going to be spending a lot of time on it. So we, regardless of what the level is, we do the most intensive review that we can. And because we know, there's a small number of people [at our organization] that do qualitative studies, and they have experience with qualitative data, so they know going in it, it's going to be a long time.

**Sara Mannheimer** 33:32

Nice. Very interesting. Okay.

**DC05** 33:56

Oh, and then to add to that, we also have a quality control process that we go through before any data sets get released. So for qualitative study, [senior staff member would] review the data set and the work that's been done and then a supervisor would release it. So there are multiple eyes on it, in case anything gets missed.

**Sara Mannheimer** 34:19

Okay. All right. Let's move to intellectual property. During your example, did you encounter challenges regarding intellectual property concerns for archiving or publishing the data? So, for example, like participant intellectual property, how do you think about that? Or intellectual property of the researchers or the programs they were studying, or anything else regarding IP? And then what strategies did you use to address them if you had challenges?

**DC05** 35:00

There are, not with this particular project, there wasn't any intellectual property concerns. The one that I, the only instance of intellectual property concerns that, that I've heard of, in my time at [the organization] would be, if questionnaires had, like copyright information in them. Like, say there's a, you know, a questionnaire scale that's been copyrighted somewhere. And they include that. I run into that a couple of times just seeing in the questionnaire, but as far as like, whose property is the qualitative data, I haven't run into that.

**Sara Mannheimer** 35:50

Okay. Alright, are there any other issues or challenges that arose during your example, or any other issues that are on your mind that we haven't talked about?

**DC05** 36:12

Um, just kind of the observation from doing this kind of work and seeing what types of funded projects like the data, the types of data that the researchers are working with. We're getting a lot more mixed methods data. So and [more PIs] have requirements on sharing their data. So there definitely needs to be more attention paid to how qualitative data gets shared, and it was, so I'm really glad that you're doing this study. Because we're seeing mixed mixed methods becoming more and more of a thing. So part of that is, we can encourage, I think there needs to be more done to encourage researchers that yes, you can share qualitative data it is possible it's being done there are places that do it. But then again, there also need to be, you know, considerations taken over, okay, if we encourage them to share their data, how are we going to take care of it? So that's important.

**Sara Mannheimer** 37:39

Have you worked with other, just out of curiosity, have you worked with like other curators who curate qualitative data? Do you, like at conferences, talk about this? Or is it just something that you develop your practices internally, mostly?

**DC05** 37:59

Um, well, I've only talked to a supervisor and mostly one other person who's had experience curating qualitative data for [my organization]. But I do know that there are like working groups and conferences, for qualitative data curation, just haven't had the opportunity to seek out those resources yet, or chat with those folks.

**Sara Mannheimer** 38:27

I feel doing qualitative research yourself is super helpful. Just knowing all of the potential pitfalls, you know, as you go through.

**DC05** 38:38

Absolutely. Especially doing disclosure risk reviews, and what to look for.

**Sara Mannheimer** 39:41

Yeah. Okay. This has been so great.

**DC05** 40:31

I'm glad. I hope it was helpful for you and the rest of your research.

**Sara Mannheimer** 40:36

Thank you. Okay.

**DC05** 40:39

Bye.